PERCEPTUAL KALMAN FILTERING SPEECH ENHANCEMENT

Chang Huai YOU[+]*, Susanto RAHARDJA*[+]*, Soo Ngee KOH**

[+] Institute for Infocomm Research, Singapore 119613
*, [+] School of EEE, Nanyang Technological University, Singapore 639798

ABSTRACT

To enhance the noisy speech signal, a perceptual weighting based Kalman filtering is investigated. Our study is to seek a high perceptual quality of speech enhancement system which optimizes the trade-off between the speech distortion and noise reduction. Using perceptual weighting to replace the masking threshold avoids the frequency domain complexity, it is suitable for time domain Kalman filtering to estimate the state-space vector in only time domain. Through many simulations, it is demonstrated that the proposed perceptual Kalman filtering outperforms the conventional Kalman filtering.

1. INTRODUCTION

Despite the wide variety of theoretically and relatively effective techniques, the problem of single channel speech enhancement still poses a challenge to the area of speech signal processing. Removing various types of noise is difficult due to the random nature of the noise and the inherent complexities of speech.

In speech enhancement area, the advantages of Kalman filtering, as compared to spectral domain processing such as spectral subtraction and MMSE spectral suppression, are that it can overcome the musical tone problem and achieve quite good speech quality to reduce the processing distortion to speech signal. Kalman filtering based speech enhancement not only exploits the statistical characteristics of signal and noise, but also utilizes the well-known speech production model, i.e., autoregressive (AR) model, which has been proven to be effective for modelling the human speech production system [1].

In Kalman filtering speech enhancement [2], one of ways to further improve the speech quality is to incorporate the perceptual properties. In [3], the time domain Kalman filtering method is extended to calculating the noise masking threshold through the Fourier transform, followed by the inverse Fourier transform with re-estimation of the speech signal based on the masking threshold. In [4], the enhanced speech from a Kalman filter is further processed by a masking-based speech enhancement system implemented in the frequency domain. However, it does not help to preserve weak spectral components which might have already been removed by the time domain Kalman filtering. In [5], the masking threshold is incorporated into Kalman filtering speech enhancement through subband analysis. In fact, the aim of the masking-based speech enhancement is to reduce the perceptual effect of the noise. It can be explained that the masking-based methods are to shape the noise spectrum with a criterion that leads the residual noise falls below the masking threshold. In [6], a perceptual weighting technique used in the speech coding [7] is incorporated into the subspace-based speech enhancement algorithm. In [8], the perceptual weighting technique is applied into the Kalman filtering speech enhancement. The speech signal itself, however, is also shaped by perceptual weighting filter; it causes an additional distortion to the speech signal during Kalman filtering.

Actually, as the Kalman filtering and AR model are processed in the time domain, the introduction of the frequency domain masking threshold into the time domain to meet the requirement of Kalman filtering processing is quite difficult. In this paper, in place of masking threshold, we propose a new method for the application of the perceptual weighting technique into time domain Kalman filtering for speech enhancement. Experiments indicate that our proposed estimator has better performance in terms of many objective measurements as compared to conventional Kalman filtering speech enhancement method. This paper is organized as follows. We introduce the Kalman filtering for speech enhancement in Section 2. In Section 3, the perceptual weighting-based Kalman filtering method is proposed. In Section 4, we show the simulation results in terms of objective assessments. Section 5 gives the conclusion and discussion for this paper.

2. KALMAN FILTER FOR SPEECH ENHANCEMENT

Consider the following model for noisy speech

$$x(n) = s(n) + v(n), \quad n = 1, 2, ... \tag{1}$$

where $x(n)$, $s(n)$ and $v(n)$ denote the discrete time samples of noisy speech, clean speech and noise respectively.

The clean speech signal is modeled as an AR process, and approximated as the output of an all-pole linear system driven by an excitation signal, $w(n)$, which is assumed to be a zero-mean white Gaussian process with variance σ_w^2, i.e.,

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + w(n). \tag{2}$$

The noise is presumed to be wide sense stationary and colored. It is adequately modelled by using the AR process [2], i.e.,

$$v(n) = \sum_{i=1}^{q} b_i v(n-i) + u(n) \tag{3}$$

where $u(n)$ is a white Gaussian process with variance σ_u^2. Let $\mathbf{s}(n) = [s(n-p+1) \quad ... \quad s(n-1) \quad s(n)]^T$, $\mathbf{v}(n) = [v(n-q+1) \quad ... \quad v(n-1) \quad v(n)]^T$, and the AR parameters $\mathbf{a}(n) = [a_p \quad ... \quad a_2 \quad a_1]^T$, $\mathbf{b} = [b_q \quad ... \quad b_2 \quad b_1]^T$. We reformulate Eqs. (1)-(3) in the form of Kalman process and measurement equations in state space domain as follows

$$\bar{\mathbf{s}}(n) = \bar{\mathbf{F}}\bar{\mathbf{s}}(n-1) + \bar{\mathbf{g}}\bar{\mathbf{w}}(n) \tag{4}$$

$$x(n) = \bar{\mathbf{C}}^T\bar{\mathbf{s}}(n) \tag{5}$$

where

$$\bar{\mathbf{s}}(n) = \begin{bmatrix}\mathbf{s}(n)\\ \mathbf{v}(n)\end{bmatrix}, \quad \bar{\mathbf{w}}(n) = \begin{bmatrix}w(n)\\ u(n)\end{bmatrix} \tag{6}$$

$$\bar{\mathbf{F}} = \begin{bmatrix}\mathbf{F} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_v\end{bmatrix}, \quad \bar{\mathbf{g}} = \begin{bmatrix}\mathbf{g} & \mathbf{0}\\ \mathbf{0} & \mathbf{g}_v\end{bmatrix}, \quad \bar{\mathbf{C}} = \begin{bmatrix}\mathbf{C}\\ \mathbf{C}_v\end{bmatrix} \tag{7}$$

with

$$\mathbf{F} = \begin{bmatrix}0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1\\ a_p & a_{p-1} & \dots & a_1\end{bmatrix}_{p\times p}, \quad \mathbf{g} = \mathbf{C} = \begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}_{p\times 1} \tag{8}$$

$$\mathbf{F}_v = \begin{bmatrix}0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1\\ b_q & b_{q-1} & \dots & b_1\end{bmatrix}_{q\times q}, \quad \mathbf{g}_v = \mathbf{C}_v = \begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}_{q\times 1}. \tag{9}$$

According to Kalman filtering theory, the estimation of state vector $\mathbf{s}(n)$ can be obtained by the following recursive equations

$$\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_n) = \hat{\bar{\mathbf{s}}}(n|\mathbf{x}_{n-1}) + \mathbf{G}(n)[x(n) - \bar{\mathbf{C}}^T\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_{n-1})] \tag{10}$$

$$\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_{n-1}) = \bar{\mathbf{F}}\hat{\bar{\mathbf{s}}}(n-1|\mathbf{x}_{n-1}) \tag{11}$$

$$\bar{\mathbf{K}}(n) = [\mathbf{I} - \mathbf{G}(n)\bar{\mathbf{C}}^T]\bar{\mathbf{K}}(n, n-1) \tag{12}$$

where

$$\mathbf{G}(n) = \bar{\mathbf{K}}(n, n-1)\bar{\mathbf{C}}[\bar{\mathbf{C}}^T\bar{\mathbf{K}}(n, n-1)\bar{\mathbf{C}}]^{-1} \tag{13}$$

$$\bar{\mathbf{K}}(n, n-1) = \bar{\mathbf{F}}\bar{\mathbf{K}}(n-1)\bar{\mathbf{F}}^T + \bar{\mathbf{Q}}. \tag{14}$$

The estimate of the speech signal is the output of Kalman filter, i.e.,

$$\hat{s}(n) = \mathbf{C}_1^T\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_n), \quad \mathbf{C}_1 = [\mathbf{C}^T \underbrace{0 \quad \dots \quad 0}_{q}]^T \tag{15}$$

where $\hat{s}(n)$ is the estimate of $s(n)$, $\mathbf{G}(n)$ is the Kalman gain, $\bar{\mathbf{K}}(n, n-1) = \mathbf{E}\{[\bar{\mathbf{s}}(n) - \hat{\bar{\mathbf{s}}}(n|\mathbf{x}_{n-1})][\bar{\mathbf{s}}(n) - \hat{\bar{\mathbf{s}}}(n|\mathbf{x}_{n-1})]^T\}$ is the predicted state-error correlation matrix. $\bar{\mathbf{Q}}$ is defined as $\bar{\mathbf{Q}} = \bar{\mathbf{g}}\mathbf{E}[\bar{\mathbf{w}}(n)\bar{\mathbf{w}}(n)^T]\bar{\mathbf{g}}^T$, which is a sparse matrix with only two nonzero elements; they are $\bar{\mathbf{Q}}(p,p) = \sigma_w^2$ and $\bar{\mathbf{Q}}(p+q, p+q) = \sigma_u^2$.

3. PERCEPTUAL SHAPING OF NOISE SPECTRUM FOR KALMAN FILTERING BASED SPEECH ENHANCEMENT

It is known that the human auditory system cannot perceive the noise within the higher-energy regions of the spectrum of a short-term speech signal, where the noise masking threshold is higher than the power spectral density of noise.

In lossy speech coding, many speech coders utility the perceptual properties of the human auditory system. Besides the masking threshold, the perceptually weighting criterion has been applied into the coding system [7]. The excitation signal is found by minimizing the weighted mean-square error over a short-term frame, where the error signal is obtained by filtering the difference between the original and the reconstructed signals through a weighting filter $W(z)$. The perceptually weighting technique trends to shape the spectra of quantization error and to limit them to be inaudible for each of the spectral regions and assures to minimize the perceptual noise caused by speech quantization while increasing the coding rate. The spectral envelop of the error can be shaped by using a suitable perceptual filter below [1]

$$W(z) = \frac{\Phi(z/\gamma_1)}{\Phi(z/\gamma_2)} = \frac{1 - \sum_{j=1}^{p} a_j\gamma_1^j z^{-j}}{1 - \sum_{j=1}^{p} a_j\gamma_2^j z^{-j}} \tag{16}$$

where $\Phi(z)$ is the linear predictive coding (LPC) polynomial, a_j is the AR coefficient of a short-term speech signal, γ_1 and γ_2 ($0 \le \gamma_2 \le \gamma_1 \le 1$) are parameters that control the energy of the error in the formant regions and p is the prediction order of the speech AR model. Fig. 1 [1] gives an example to show the frequency response of the perceptual weighting filter for a speech segment (32 ms) with different settings of the γ_1 and γ_2.

As like conventional Kalman filtering, we have to estimate the AR coefficients for both speech and noise models. The time-varying parameters $\mathbf{a} = [a_p \quad ... \quad a_2 \quad a_1]^T$ can be computed according to the Levison-Durbin algorithm. Subsequently, the variance of the excitation of the speech model, σ_w^2, can be computed by

$$\begin{aligned}\sigma_w^2 &= \mathbf{E}\{|s(n) - \sum_{i=1}^{p} a_i^*(n)s(n-i)|^2\}\\ &= R_{ss}(0) - 2Re\{\sum_{i=1}^{p} a_i^* R_{ss}(i)\} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i^* a_j R_{ss}(i-j)\end{aligned} \tag{17}$$

here R_{ss} denotes the autocorrelation of the speech signal.

Based on the AR model of the (pre-estimated) speech signal, we shape the spectrum of noise by applying the perceptual weighting filter in the time domain, so that a perceptually

[1] In this figure, the dB number given to the perceptual weighting filtering is the relative values as for reference only to show the spectral envelopes in contrast with the signal spectral envelop.

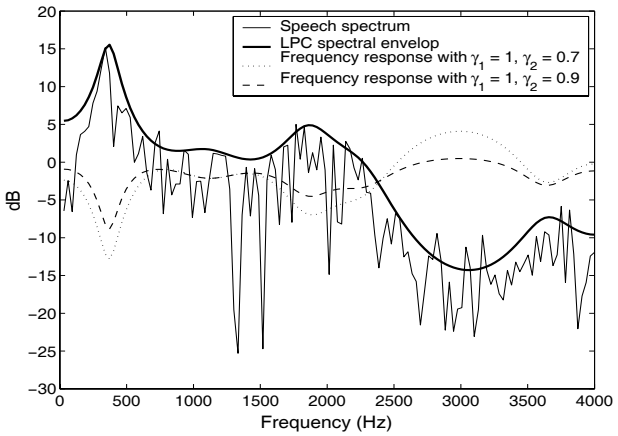

Fig. 1. The spectral amplitude of a segment of a speech signal, its LPC spectral envelop with $p = 10$, and the frequency response of the corresponding perceptual weight filter with $\gamma_1 = 1, \gamma_2 = 0.7$ and $\gamma_1 = 1, \gamma_2 = 0.9$ respectively.

shaped noise AR model is estimated, we have

$$v(n) = \sum_{k=1}^{p} a_k \gamma_2^k v(n-k) + v_o(n) - \sum_{k=1}^{p} a_k \gamma_1^k v_o(n-k) \quad (18)$$

where $v_o(n)$ denotes the original estimated noise signal, $v(n)$ is the shaped noise via the perceptual weighting filter.

Using Kalman filtering to estimate the speech signal, generally we have to estimate the AR parameters of noise signal. In contrary with the conventional colored noise model Kalman filtering where AR parameters of noise signal is directly obtained from the estimated noise signal, we propose that the AR parameters of the noise signal [2] is obtained from the shaped noise signal $v(n)$, which is the modified version of noise by applying the perceptual weighting filter to the original estimated noise signal, $v_o(n)$. In other words, $\mathbf{b} = [b_q \quad ... \quad b_2 \quad b_1]^T$ are computed according to the Levison-Durbin algorithm from the shaped noise $v(n)$, and σ_u^2 is obtained as follows

$$\begin{aligned} \sigma_u^2 &= \mathbf{E}\{|v(n) - \sum_{i=1}^{q} b_i^*(n) v(n-i)|^2\} \\ &= R_{nn}(0) - 2Re\{\sum_{i=1}^{q} b_i^* R_{nn}(i)\} + \sum_{i=1}^{q}\sum_{j=1}^{q} b_i^* b_j R_{nn}(i-j) \end{aligned} \quad (19)$$

where R_{nn} represents the autocorrelation of the perceptually shaped noise. Fig. 2 describes the procedure of our proposed perceptual Kalman filtering speech enhancement system.

[2]White noise can be considered as a special case of the colored noise model. In general, even if the original noise is white, the shaped noise becomes colored.

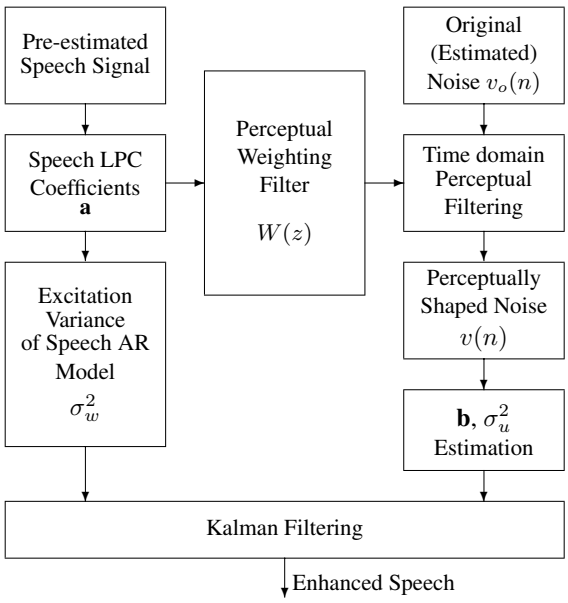

Fig. 2. The flow chart of the Kalman filtering speech enhancement system with the perceptual noise shaping.

4. PERFORMANCE EVALUATION

In the simulation, we use twenty speech utterances from the TIMIT database for the evaluation. Half of the utterances are produced by male speakers and half by female. The effectiveness of the enhancement algorithms is evaluated at the sampling rate of 8 kHz which is down-sampled from 16 kHz after pre-filtering.

For 8 kHz sampling rate, the AR order of speech model is set to $p = 10$. In order to accurately represent the shaped noise spectral envelop, we select $q = 5$ for noise AR order. We use 32 ms (256 samples for 8kHz sampling rate) for autocorrelation computation. The current time step locates at the central point of the 32 ms range. For each 5 ms, the AR coefficients **a** and **b** are computed through the Levinson-Durbin algorithm. For each time step of the Kalman filtering, σ_w^2 and σ_u^2 are updated once. Consequently the total time delay of the Kalman filtering system is 16 ms.

For the sake of predicting the best and worst performance of our proposed method, we consider two cases to estimate the AR parameters of speech signal. One extreme case is that the time-varying AR parameters of the speech model are estimated using the clean speech, we call it as the "Ideal-AR" case. Another extreme case is that the AR parameters of speech signal are estimated using the noisy speech, which we shall call the "Noisy-AR" case. Table 1 shows the performances of the conventional Kalman filtering and the proposed perceptual Kalman filtering with $\gamma_1 = 0.9$ and (1) PerKal1: $\gamma_2 = 0.9$, (2) PerKal2: $\gamma_2 = 0.8$, (3) PerKal3: $\gamma_2 = 0.7$, in terms of segmental SNR (seg.SNR), Itakura-Saito (IS) distortion and modified Bark spectral distortion (MBSD) measures, where the speech utterances are contaminated by white noise

Table 1. Comparison of the performances between the conventional Kalman filtering and the proposed perceptual Kalman filtering, where the input seg.SNR is -5 dB.

	Methods	Objective Measurements		
		seg.SNR	IS	MBSD
Ideal -AR	StdKal	**5.65**	0.46	0.89
	PerKal1	5.62	0.33	0.62
	PerKal2	5.57	0.30	0.59
	PerKal3	5.50	**0.29**	**0.57**
Noisy -AR	StdKal	-1.96	0.68	2.65
	PerKal1	-1.65	0.54	2.20
	PerKal2	-1.60	0.53	2.16
	PerKal3	**-1.57**	**0.52**	**2.14**

Table 2. Comparison of the performances between the conventional Kalman filtering and the proposed perceptual Kalman filtering, where the input seg.SNR is 0 dB.

	Methods	Objective Measurements		
		seg.SNR	IS	MBSD
Ideal -AR	StdKal	**8.20**	0.26	0.52
	PerKal1	8.16	0.21	0.42
	PerKal2	8.11	0.19	0.39
	PerKal3	8.04	**0.18**	**0.36**
Noisy -AR	StdKal	2.06	0.37	1.50
	PerKal1	2.71	0.29	1.26
	PerKal2	2.75	0.28	1.24
	PerKal3	**2.78**	**0.27**	**1.22**

with the average seg.SNR = -5 dB, IS = 0.84 and MBSD = 4.00. Table 2 is under the same simulation conditions as Table 1 except the input noisy speech utterances are with the average seg.SNR = 0 dB, IS = 0.45 and MBSD = 2.26.

From the simulation results [3], we can see that the performance of the proposed method is generally better than the conventional Kalman filtering except the seg.SNR measure in the Ideal-AR case; and the performance is better for the proposed method when the weighting degree increases with γ_2 value from 0.9 to 0.7.

5. CONCLUSION AND DISCUSSION

In this paper, we propose a new method for the application of the perceptual weighting filtering into the Kalman filtering-based speech enhancement. Through computer simulations, we investigate the degrees of the effect of the perceptual weighting filtering along with the weighting factors (γ_1 and γ_2) as well as the comparison with the conventional Kalman filtering, in terms of seg.SNR, IS distortion and MBSD measures. The simulation results show that the proposed method is generally effective and improved over the conventional Kalman filtering based speech enhancement method. Although the results provided here are only for the case of white noise, in our simulation, the similar findings are also suitable for the cases of F-16 cockpit noise and Volvo car interior noise.

Here we compute the AR parameters in the two extreme cases (i.e., Ideal-AR and Noisy-AR) for the Kalman filtering method. One may evaluate the performances for the various iterations of the enhanced speech obtained by the Kalman filtering, i.e., the enhanced speech signal can be obtained through Kalman filtering by using the AR parameters obtained from the estimated speech signal in the previous iteration of the Kalman filtering, while the original noisy signal is always considered as input signal to the Kalman filter.

[3] The value with bold style represents the best one.

6. REFERENCES

[1] Sadaoki Furui and M. Mohan Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker, Inc., NY, 1992.

[2] J.D. Gibson, B. Koo, and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1732-1742, Aug.1991.

[3] N. Ma, M. Bouchard and R. A. Goubran, "A Perceptual Kalman Filtering-Based Approach for Speech Enhancement," *International Symposium for Signal Processing and Applications*, ISSPA, Jul. 2003.

[4] N. Ma, M. Bouchard and R. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-2004, Vol. 1, pp. 717-720, May 2004.

[5] C.H. You, S.N. Koh and S. Rahardja, "Kalman Filtering Speech Enhancement Incorporating Masking Properties for Mobile Communication in a Car Environment," *Proc. IEEE International Conference on Multimedia and Expo*, ICME'2004, Taiwan, Jun. 2004.

[6] Y. Hu and P.C. Loizou, "A Perceptually Motivated Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 5, pp 457-465, Sep. 2003.

[7] P. Kroon and B.S. Atal, "Predictive Coding of Speech Using Analysis-by-Synthesis techniques," in Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., pp. 141-164. Marcel Dekker, NY, 1992.

[8] V. Grancharov, J. Samuelsson and W.B. Kleijn, "Improved Kalman Filtering for Speech Enhhancement", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-2005, Vol. 1, pp. 1109-1112, Mar. 2005.